MSU Deep Learning Course

Image Captioning

Ashuha Arseniy[1,2]

Bayesian Research Group[1], MIPT[2]

ars-ashuha.ru/slides

November 25, 2016

Motivation

- What do we want?

"man in black shirt is playing guitar."

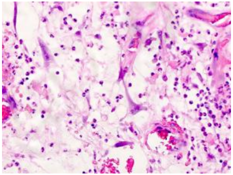
"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

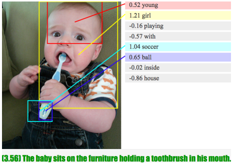
"boy is doing backflip on wakeboard."

- Applications

(a) Your's example



(b) Medicine

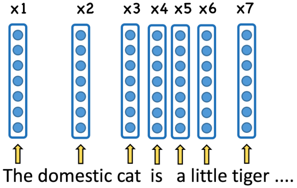


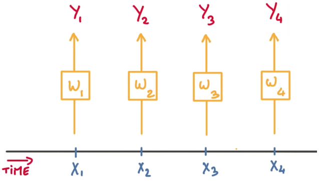
(c) Search engine

http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

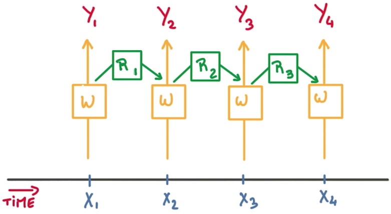
Recurrent Nets Recap

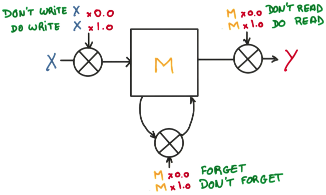
- We want to process a number of inputs, each item is a sequence





- We will add recurrent connections and memory control options





- `lasagne.layers.EmbeddingLayer, lasagne.layers.LSTMLayer`

Text Generative Model

- What is a text generative model?

$$p(w_{n+1}|w_1, \dots, w_n; \theta)$$

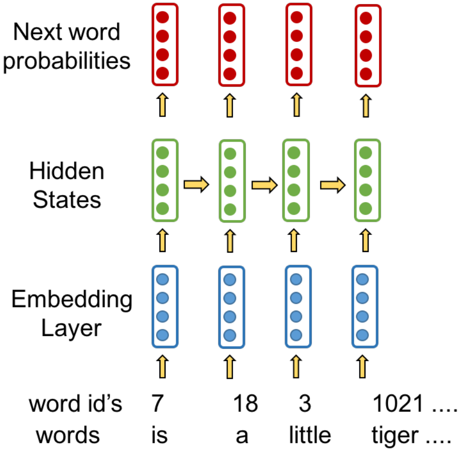
- Let's specify dimensions:
  - $p(w_{n+1}).shape$ is vocabulary size
  - $w_i.shape$ is embedding size
  - $\theta.shape$ has shape like parameters
- How can we define this function? RNN!

$$p(w_{n+1}|w_1, \dots, w_n; \theta) = p(w_{n+1}|w_n, hiden_{n-1}; \theta)$$

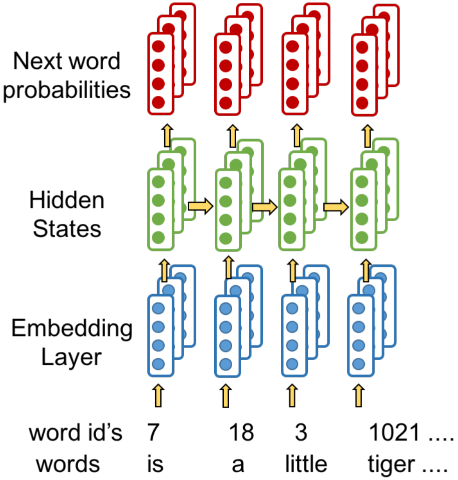
But, there are several technical problems to apply it

- How would you to construct a batch?
- How to define loss functions?
- How to calculate soft-max over millions point? ()

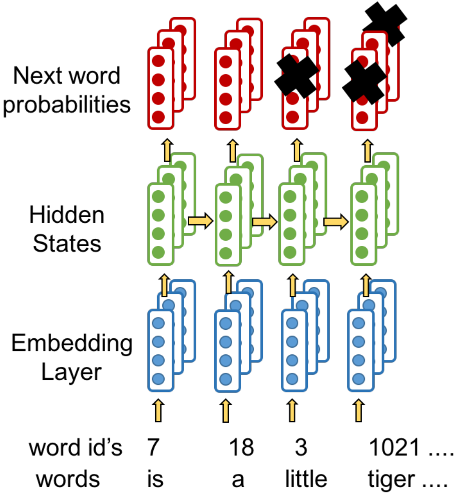
Tech Details and General Scheme



Tech Details and General Scheme

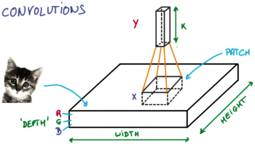


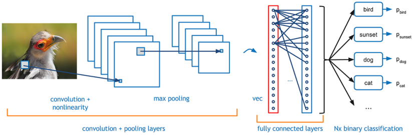
Tech Details and General Scheme



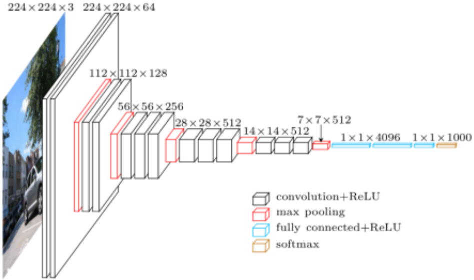
Convolution Net Recap

- Convolution (lower number of parameters cw fully connected)
- Convolution and Convolution Neural Net



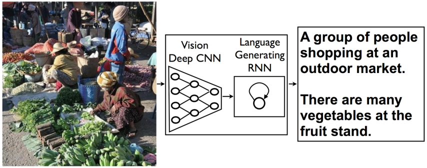


- Morned Architecture (VGG16)



Condition Text Generative Model

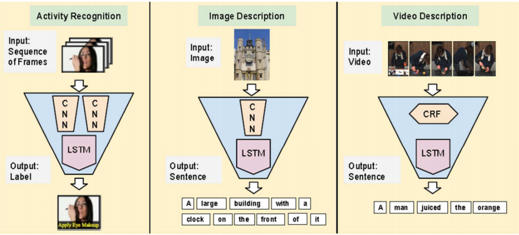
- How to build image captioning Neural Nets?



A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

- Another useful thinks?



A large building with a clock on the front of it

A man juiced the orange

But, extremely complex to train

http://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Code should be like this

```
sentences = T.imatrix()# word ids
image_vec = T.matrix() # image features
sentence_mask = T.neq(sentences, PAD_ix)

words = InputLayer((None, None), sentences)
masks = InputLayer((None, None), sentence_mask)
words_emb = EmbeddingLayer(words, n_tokens, EMBED_SIZE)

image_fea = InputLayer((None, CNN_FEATURE_SIZE), image_vec)
image_emb = DropoutLayer(image_fea, 0.5)
image_emb = DenseLayer(image_emb, LSTM_UNITS)

decoder = LSTMLayer(
        words_emb, num_units=LSTM_UNITS,
        cell_init=image_emb, mask_input=mask)

decoder = Reshape (...)
loss = categorical_cross_enrtopy (...)
```

AWS Educate

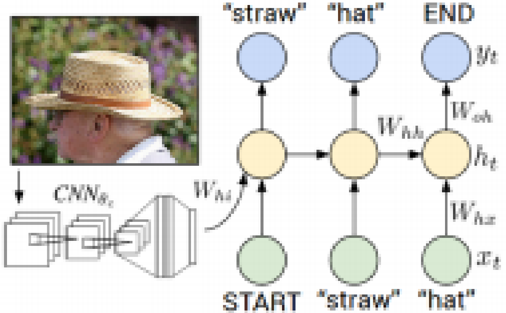
- Training is extremely hard, so you should use pre-train models
- If it's still hard, let's use gpu

- But gpu is a little bit expensive
- AWS Educate give 100$ grand for all MSU students ≈ 160 hours
- You need only MSU email
- You can rent GPU server on Amazon WS free using this money!!!

https://aws.amazon.com/ru/education/awseducate/

Go To Break

Image Captioning Seminar



http://mybinder.org/repo/ars-ashuha/caption_binder